	BioTeam, Inc. http://www.bioteam.net **Primary Contact:** Bhanu Rekepalli, Ph.D., Senior Director of Government Services **Email:** bhanu@bioteam.net **Tel:** 865-230-1605 Contract Number: 47QTCA19D00JP Order Number: 75N97022F00028

NIH/NLM: Root Cause Analysis: Removal of SRA Sequence Data - SUMMARY

Prepared for: Dr. Stephen Sherry, Acting Director, NLM/NCBI

Prepared by: Ari Berman, Ph.D., CEO, BioTeam
Laura Boykin, Ph.D., Senior Scientific Consultant, BioTeam
Anna Sowa, Ph.D., Senior Scientific Consultant, BioTeam
Simon Twigger, Ph.D. Director of Data Science, BioTeam
Myra Ceasar, Senior Delivery Services Consultant, BioTeam

Document History: Summary delivered March 11th, 2022.

Summary of BioTeam Report: "NIH/NLM: Root Cause Analysis: Removal of SRA Sequence Data"

Background: The Sequence Read Archive (SRA) at the National Institutes of Health (NIH), National Library of Medicine (NLM) is the largest publicly available repository for high throughput genetic sequence (nucleotide) data in the world. It contains genetic data from many forms of life on Earth and are deposited by scientists worldwide. NLM's mission is to collect, organize, and disseminate scientific information relevant to the progress of medicine and public health. SRA aims to maintain a permanent record of sequence data generated using high-throughput sequencing methods.

Since the pandemic began, SRA has been a major destination for global genomic sequences for the virus that causes COVID-19 (SARS-CoV-2). On March 16th, 2020, 241[1] SARS-CoV-2 genetic sequences from samples collected in Wuhan, China, were submitted to SRA by a researcher from Wuhan University. The sequences were made public at first but were later removed from SRA public access at the researcher's request on June 17th, 2020. This public summary provides a high-level overview of issues that arose surrounding those SARS-CoV-2 sequences, along with causes, and outcomes from those events.

On June 21, 2021, a pre-publication scientific paper (not peer-reviewed)[2] was submitted to the online BioRxiv website[3], which referenced these same 241 sequences. The paper noted that the sequence data had been made public and then were suddenly removed from SRA. The paper also documented that, while the sequences were no longer accessible via the SRA website, the data could still be accessed in an additional SRA repository in a public cloud. The paper implied that these sequences from the beginning of the pandemic could hold key information about the origin of COVID-19 and may have been mishandled by SRA.

Subsequently, NLM initiated an analysis to understand how SRA had managed these 241 sequences. The primary goal of this analysis was to determine if the sequences had been handled in a manner consistent with SRA policies and procedures. The secondary goal was to identify areas of opportunity for SRA to improve its processes going forward. Consistent with industry norms, the analysis was undertaken by an independent third party (The BioTeam, Inc. (https://bioteam.net) between July and August 2021.

Findings: Two sets of policies define how SRA should handle submitted sequence data. The first are NLM's policies, which govern its approach to collection and preservation of materials.

[1] One additional sequence was submitted by the researcher on June 5, 2020 yielding an overall total of 242 sequences.
[2] A fully peer reviewed version of this article was published on 8/16/2021 - https://doi.org/10.1093/molbev/msab246
[3] 'Recovery of deleted deep sequencing data sheds more light on the early Wuhan SARS-CoV-2 epidemic', Jesse D. Bloom https://www.biorxiv.org/content/10.1101/2021.06.18.449051v2

The second are INSDC[4,5] policies, which describe how INSDC members like NLM should manage sequence data. The independent analysis identified three specific instances where sequence data handling in the case of the 241 SARS-CoV-2 sequences from Wuhan, China, did not conform to the standard policies of both NLM and INSDC:

1. **SRA staff used the wrong workflow to meet the Wuhan scientist's request to remove the sequences from the public database**. The SRA database has two states for submitted sequence data: unpublished and public. Unpublished sequences are not available to the public through the SRA website or tools. Anyone can access public sequence data and these data form part of the permanent scientific record. The Wuhan sequences were public from March 16th, 2020 until June 17th, 2020 when, at the request of the scientist that submitted these data, they were removed from public access. Per INSDC policy [5], public sequence data can be removed based on a submitter request if data are submitted in error or if there is a quality issue[5]. Data removed this way remain publicly accessible based on the sequence identifier but are not found using general search terms. SRA database managers can also return data to an unpublished status if the database manager has erroneously released data prematurely, or if data were submitted without the permission of the rightful owner[5]. *In the case of the Wuhan data, instead of using the command for removing public data per submitter request, the command to return data to the unpublished state was used in error.* This action resulted in the sequences being removed but also made them inaccessible via their original accession numbers. This result was incorrect because the sequences were already public, and the policy is to retain public access by accession numbers.

2. **The processes that SRA used to ensure the sequences were correctly removed from all systems did not synchronize with their cloud storage locations**. SRA maintains a dual-storage system, in a local datacenter at NLM, and in public cloud environments. The sequences should not have been accessible in either SRA location after the incorrect command to return the sequences to the 'unpublished' status was run. However, they were still found in the cloud environment through the sequence identifier[3]. This inconsistency highlighted a gap in SRA's system management that was not consistent with NLM's and INSDC's policies.

3. **All 241 sequences remained available in the SRA system in SRAs normalized data format, however, 18 of the 241 original data files were inadvertently deleted.** In the first half of 2020, the SRA team was undergoing considerable growth fueled by the increase in volume submissions due to the COVID-19 pandemic. At the same time, there were efforts underway to move sequence data to the cloud. A detailed analysis of file management efforts revealed that, while SRA's sequence data files created from the original submission existed for all of the Wuhan sequences, 18 of the 241 originally submitted sequence files were not transferred to the cloud before local copies of the files were removed to clear disk space. This loss of 18 of the original 241 sequences did not meet NLM's goal of preserving its collected data.

[4] International Nucleotide Sequence Database Consortium (INSDC), https://www.insdc.org/
[5] INSDC Status Document, https://www.insdc.org/documents/insdc-status-document

Contributing factors and future considerations: The analysis identified a few key factors that contributed to each of the findings above:

For Issue 1 above, the following factors were identified:

- NLM staff working on SRA did not manage this specific removal request according to SRA's generally accepted practices.
- SRA's existing policies and procedures did not sufficiently document, and the staff was not adequately trained on the process for submission withdrawal.
- There was no official escalation or review process in place to validate potentially high-impact actions.
- SRA's practices and procedures did not have provisions for extraordinary situations (such as a global pandemic) that may require exceptions to standard policies and procedures.
- SRA's existing policies, procedures and staff management processes lagged behind the urgent efforts to expand compute resources to handle the ever-increasing demand for managing genomic sequences
- There were no technical fail-safes to prevent SRA staff from taking an inappropriate action on public sequence data.

For Issue 2, the following factors were identified:

- Automated processes for keeping cloud storage in sync with the main SRA system had not yet been implemented at the time in question. The use of cloud storage for sequence data was a multi-year planned process to migrate 35 petabytes of data from local storage to public clouds that were greatly accelerated due to the pandemic. Before 2019, sequence data was only stored in NLM's local datacenter. The accelerated file movement was performed manually and rapidly before automated checks and balances could be implemented by SRA.
- Managing sequence data storage and access in the cloud requires different processes than managing the same files in a local system. No policies have been formulated to guide managing public sequence data in the cloud.

For Issue 3, the following factors were identified:

- The rapid and unexpected influx of pandemic data stressed SRA's local disk capacity, and it was close to running out of space at the time in question. An existing initiative to migrate to cloud storage (mentioned above) had to be greatly accelerated in March-April of 2020 to keep SRA systems functioning.
- As part of the accelerated data migration, much of the sequence data file movement to the cloud was initially done manually and lacked oversight. Unfortunately, undetected errors during these file movements led to a small number of sequences in original

submission format not being copied to the cloud. This data was subsequently lost when the local data center copies were deleted to save disk space.

A variety of **systemic factors** also contributed to the issues described above:

- SRA prioritized customer-facing activities to meet the immediate needs of the scientists over the need to develop and maintain internal policies and procedures, train staff, and make improvements to internal sequence data management systems and tools.
- NLM's strategy for SRA to move data and systems to the cloud deliberately emphasized the technical engineering of the SRA system for durability rather than investing in local data center computational infrastructure. This plan was intended to reduce costs and take advantage of the cloud and the NIH STRIDES[6] program.
- Overall budgetary constraints - SRA continues to observe data growth, but the program budget remains flat. The financial constraints impact SRA's activities around hiring and training, maintenance of infrastructure (providing enough storage space for the data), and the creation and maintenance of internal systems.
- The COVID-19 pandemic added to the general sense of urgency for data handling at SRA. NIH's policies governing SRA during such a public health emergency are not clearly defined and represent a significant departure from the research-centric sequence submissions and inquiries SRA handles typically.

Identified Opportunities: The independent analysis identified the following key opportunities to address the contributory factors:

- Re-release the 241 sequences in question in SRA (make them public again) since they have been published by the original researcher.
- Strengthen cloud storage and management policies for NLM and SRA policy for SRA.
- Make improvements to internal policies and procedures, including training.
- Work with NIH to align policies and investment supporting the role of SRA in public health emergencies and its data management responsibilities.
- Modernize the internal command interfaces to improve efficiency and minimize the risk of inadvertent errors.

[6]NIH Science and Technology Research Infrastructure for Discovery, Experimentation, and Sustainability (STRIDES) Initiative - https://datascience.nih.gov/strides.